Learning Event-Specific Localization Preferences for Audio-Visual Event Localization

Shiping Ge
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
shipingge@smail.nju.edu.cn

Zhiwei Jiang*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
jzw@nju.edu.cn

Yafeng Yin
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
yafeng@nju.edu.cn

Cong Wang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
cw@smail.nju.edu.cn

Zifeng Cheng
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
chengzf@smail.nju.edu.cn

Qing Gu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
guq@nju.edu.cn

ABSTRACT

Audio-Visual Event Localization (AVEL) aims to locate events that are both visible and audible in a video. Existing AVEL methods primarily focus on learning generic localization patterns that are applicable to all events. However, events often exhibit modality biases, such as visual-dominated, audio-dominated, or modality-balanced, which can lead to different localization preferences. These preferences may be overlooked by existing methods, resulting in unsatisfactory localization performance. To address this issue, this paper proposes a novel event-aware localization paradigm, which first identifies the event category and then leverages localization preferences specific to that event for improved event localization. To achieve this, we introduce a memory-assisted metric learning framework, which utilizes historic segments as anchors to adjust the unified representation space for both event classification and event localization. To provide sufficient information for this metric learning, we design a spatial-temporal audio-visual fusion encoder to capture the spatial and temporal interaction between audio and visual modalities. Extensive experiments on the public AVE dataset in both fully-supervised and weakly-supervised settings demonstrate the effectiveness of our approach. Code will be released at https://github.com/ShipingGe/AVEL.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Scene understanding**.

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612506

KEYWORDS

Audio-Visual Event Localization, Memory Bank, Contrastive Learning

ACM Reference Format:
Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. 2023. Learning Event-Specific Localization Preferences for Audio-Visual Event Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612506

1 INTRODUCTION

Audio-Visual Event Localization (AVEL) is a challenging task that has received growing attention in recent years [22]. Its objective is to identify the category of events in a video while simultaneously locating the video segment where the event occurs to be both audible and visible.

Existing AVEL approaches primarily consist of two mainstream paradigms: single-branch paradigm [19, 22, 32, 34], and double-branch paradigm [27–29, 31]. Among them, as shown in Figure 1(a-b), the single-branch paradigm considers the problem as an (N+1) classification problem at the segment level, including N event classes and one background class. The double-branch paradigm decomposes the problem into an N-class event classification problem at the video level, and a two-class event localization problem at the segment level. Despite the effectiveness of these two paradigms, they both set a general background class for event localization, aiming to learn generic localization patterns that are applicable to all events. However, they are prone to overlooking the localization preferences specific to different events.

Actually, different types of events often have their own special event localization preferences due to the modality bias of the event. As shown in Figure 1(d-f), for some events, the visual aspect is relatively more important, such as "airplane" events where the noise is constant throughout and it is difficult to extract localization information from audio, and the visual aspect often dominates the localization. For some events, the audio aspect is relatively more important, such as "man speaking" events where there is less visual variation and it is difficult to extract localization information from

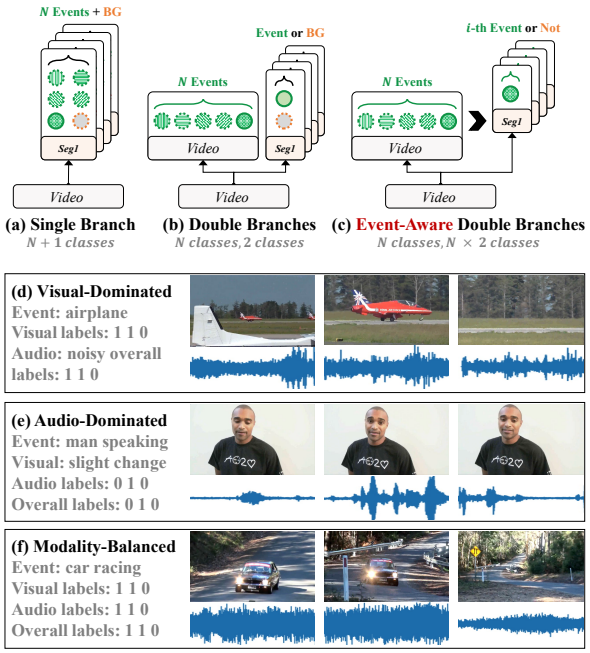


Figure 1: Illustration of three AVEL paradigms (a-c) and three types of events with different localization preferences (d-f).

the visual aspect, and the audio aspect often dominates the localization. There are also events where the visual and audio aspects are balanced, such as "car racing" events where both modalities provide relatively clear localization information and it is easier to make comprehensive decisions using both modalities. Capturing these event-specific localization preferences may help enhance event localization performance.

To address this issue, we propose a novel event-aware double-branch localization paradigm, which first identifies the event category of the video and then utilizes the localization preferences specific to that category to perform segment-level event localization, as shown in Figure 1(c). Compared to the previous two paradigms, this event localization paradigm can more fully exploit the localization patterns possessed by videos with the same event category, thus achieving better localization performance.

To achieve this, we propose a memory-assisted metric learning framework that can store segments with high event-relatedness and high audio-visual synchronization in the memory during the training process for each event category, and then use these segments as anchors to adjust the distribution of the representation space. In particular, we learn a unified representation space that can simultaneously perform event classification and event localization, which not only benefits both tasks with each other but also avoids the inconsistencies that may arise from learning separate models for the two tasks.

Although the above metric learning framework is not designed specifically for any particular model structure, it places high demands on the ability of the model to extract multimodal representations. It requires the extracted segment representations to capture some subtle changes, such as the temporal-spatial variations in the visual modality and the temporal variations in the presence of noisy audio. To this end, we propose a spatial-temporal audio-visual fusion encoder to capture the spatial-temporal interaction between visual and audio modalities.

The contributions of this paper are as follows:

- We propose a new event-aware double-branch localization paradigm to utilize event preferences for more accurate localization.
- We propose a spatial-temporal audio-visual fusion encoder and a memory-assisted metric learning framework to better extract features and capture event-specific localization preferences.
- Extensive experiments on the public AVE dataset in both fully-supervised and weakly-supervised settings demonstrate the effectiveness of our approach.

2 RELATED WORK

2.1 Audio-Visual Event Localization

Audio-Visual Event Localization (AVEL) is a significant task with a diverse range of applications, such as video surveillance, human-robot interaction, and multimedia content analysis [22]. Existing methods primarily focus on designing effective model architectures that can extract cross-modal relations between audio and visual features, utilizing co-attention fusion [19, 22, 26–28, 31, 34] or transformer fusion [13, 17, 29, 32]. In terms of co-attention fusion, Tian et al. [22] first introduce the AVEL problem and collect an Audio-Visual Event (AVE) dataset, which has become the standard benchmark for evaluating AVEL methods. Wu et al. [27] propose a Dual Attention Matching (DAM) module to cover a longer video duration for better high-level event information modeling. Ramaswamy [19] propose a novel Audio-Visual Interacting Network (AVIN) that enables inter as well as intra-modality interactions. Yu et al. [31] propose a Multimodal Parallel Network (MPN) to perceive global semantics and unmixed local information parallelly. Zhou et al. [34] propose a Positive Sample Propagation (PSP) module to discover and exploit the closely related audio-visual pairs. Xia and Zhao [28] proposes a novel Cross-Modal Background Suppression (CMMS) network to improve localization performance by suppressing asynchronous audio-visual background frames. Wu et al. [26] propose a span-based framework that considers consecutive segments jointly. As for transformer fusion, Lin and Wang [13] propose an Audio-Visual Transformer (AVT) to exploit intra and inter-frame visual information and perform co-attention over different modalities. Xu et al. [29] propose a Cross-Modal Relation-Aware Network (CM-RAN) which contains an audio-guided attention module to guide the model to focus on event-relevant visual regions. Yu et al. [32] propose a Multi-Modal Pyramid Attentional Network (MM-Pyramid) to capture temporal pyramid features and integrate pyramid features interactively with an adaptive semantic fusion module. Mahmud and Marculescu [17] integrate the AudioCLIP model [5] pre-trained on large-scale audio-visual data to effectively operate on different temporal scales of video frames.

2.2 Memory Networks

Memory networks are a class of learning models [25] that construct an external memory bank module to store potentially useful features for future use [8]. Pioneering work using memory networks has shown great potential in NLP and CV tasks [1]. Zeng

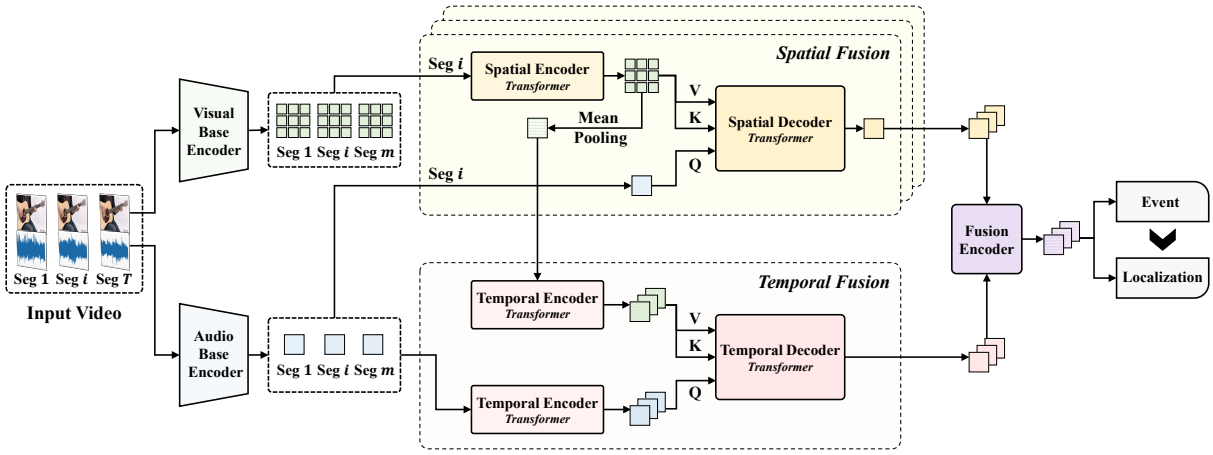


Figure 2: The illustration of spatial-temporal audio-visual fusion encoder and the event-aware double branches.

et al. [33] propose topic memory networks for short text classification with a novel topic memory mechanism to encode latent topic representations indicative of class labels. Liu et al. [14] propose a memory-guided semantic learning network to record the shared semantic features in the temporal sentence grounding task. Kim et al. [9] propose a novel memory-guided domain generalization method for semantic segmentation by learning how to memorize domain-agnostic and distinct information of classes. Ji and Yao [8] design a memory module to memorize the blurry-sharp feature pairs in the memory bank, thus providing useful information for video deblurring. Recently, some researchers have applied memory networks to multi-modal learning problems. Liu et al. [15] propose a Memory-Augmented Unidirectional Metric Learning Method to enhance the cross-modality association by storing the modality-specific proxies into memory banks to increase the reference diversity. Chen et al. [2] propose a cross-modal memory network to enhance the encoder-decoder framework for radiology report generation, where shared memory is designed to record the alignment between images and texts. Li and Moens [12] propose a memory-enhanced graph network that performs explicit and implicit reasoning over a key-value knowledge memory module for visual question answering.

3 METHODOLOGY

3.1 Task Definition

We first introduce notations and formalize the Audio-Visual Event Localization (AVEL) task. For the AVEL task, each video V is split into T non-overlapping segments S, i.e., $V = \{S_t\}_{t=1}^{T}$ and $S_t = (v_t, a_t)$, where v_t and a_t are the visual and audio feature of the segment S_t, respectively. Let $Y = \{\{y_t^n | y_t^n \in \{0,1\}\}_{n=1}^{N+1}, \sum_{n=1}^{N+1} y_t^n = 1\}_{t=1}^{T}$ represent the event label for video V. Here, $N+1$ denotes the number of event categories, including a background category indicating independently audible (or visible) events or the absence of an event [20, 22]. Since one video usually contains only one event, the event labels y can be further decomposed into two sub-labels: (1) video-level event category label $y^e \in \{1, \ldots, N\}$, (2) segment-level event relevance label $y^r = \{y_t^r | y_t^r \in \{0,1\}\}_{t=1}^{T}$.

We consider two settings of the AVEL task: 1) **Fully-supervised AVEL** setting, in which the event label y_t of each video segment S_t is given during training. 2) **Weakly-supervised AVEL** setting, in which only the video-level event category label y^e for the whole video V is given during training. The goal of AVEL is to predict the event label Y_t for each video segment S_t in both fully-supervised and weakly-supervised AVEL settings.

3.2 Overview

To address the AVEL task, we propose a model structured with a spatial-temporal audio-visual fusion encoder for representation and a novel Event-Aware Double Branches (EADB) for inference, as shown in Figure 2. Moreover, to effectively learn event-specific localization preferences, we develop a Memory-Assisted Metric Learning (MAML) framework. Specifically, the model takes a sequence of video segments as inputs and extracts features for further event localization by both spatial and temporal audio-visual feature fusion operations. These features are finally sent to the EADB for both event classification and localization. The MAML framework is designed based on the memory bank mechanism, which allows us to store segments with high event-relatedness and high audio-visual synchronization in the memory for each event category. These stored segments can then be used as anchors to adjust the distribution of the representation space based on metric learning.

3.3 Spatial-Temporal Audio-Visual Fusion Encoder

As shown in Figure 2, we begin by feeding the raw visual feature maps and raw audio features into the base visual encoder E_b^v and the base audio encoder E_b^a, respectively, to obtain the visual feature maps $F^v = \{f_1^v, \ldots, f_T^v\} \in \mathbb{R}^{T \times h \times w \times d_v}$ and the audio feature $F^a = \{f_1^a, \ldots, f_T^a\} \in \mathbb{R}^{T \times d_a}$. Here, T is the number of segments, d_v represents the number of channels of the visual feature, d_a represents the number of channels of the audio feature, and h and w denote the height and width of the visual feature map, respectively.

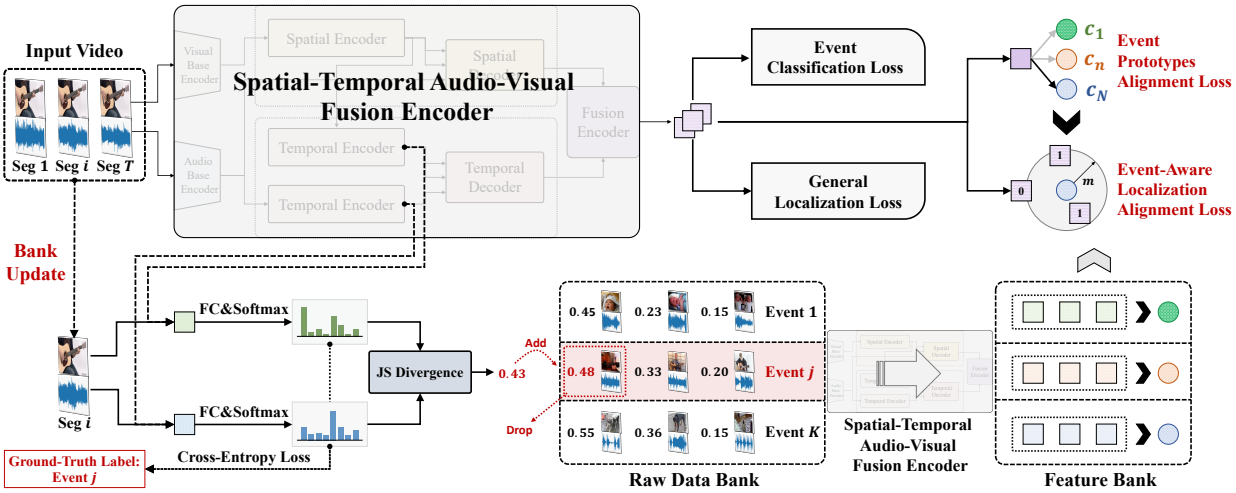


Figure 3: The illustration of our proposed memory-assisted metric learning framework.

3.3.1 ***Spatial Audio-Visual Feature Fusion.*** In this module, we first flatten each visual feature map $f_i^v \in \mathbb{R}^{h\times w\times d}$ into a sequential representation $\tilde{f}_i^v \in \mathbb{R}^{hw\times d}$, which is then input into a transformer encoder E_s^v to obtain the spatial context-aware visual feature f_s^v. Subsequently, we feed each audio feature $f_i^a \in F^a$ and the corresponding f_{si}^v into a transformer decoder D_s, producing the spatially fused feature f_{si}^{av}. We represent the entire sequence of spatially fused features as $F_s^{av} = \{f_{s1}^{av}, \ldots, f_{sT}^{av}\}$.

3.3.2 ***Temporal Audio-Visual Feature Fusion.*** In this module, we begin by inputting the audio feature F^a into a transformer encoder E_t^a, resulting in the temporal context-aware audio feature $F_t^a \in \mathbb{R}^{T\times d}$. Next, we compute the mean of each spatial context-aware visual feature map $\tilde{f}_i^v \in \mathbb{R}^{hw\times d}$ to obtain the visual feature $\bar{f}_i^v \in \mathbb{R}^d$. The set of all visual features is represented as $\bar{F}^v = \{\bar{f}_1^v, \ldots, \bar{f}_T^v\} \in \mathbb{R}^{T\times d}$. We then input $\bar{F}^v$ into a transformer encoder E_t^v to obtain the temporal context-aware video feature $\bar{F}_t^v \in \mathbb{R}^{T\times d}$. Finally, we input F_t^a and $\bar{F}_t^v$ into a transformer decoder D_t, resulting in the temporal fused features $F_t^{av} = \{f_{t1}^{av}, \ldots, f_{tT}^{av}\}$.

3.3.3 ***Final Fusion.*** In the final step, we concatenate the spatial and temporal features of each segment and feed them into a convolutional network to obtain the final fused features, denoted by F^{av}. Specifically, we apply a 1D convolution operation with a ReLU activation function to the concatenation of each pair of corresponding spatial and temporal features:

$$F^{av} = \text{Conv}(\text{Concat}[f_{s1}^{av}, f_{t1}^{av}], \ldots, \text{Concat}[f_{sT}^{av}, \ldots, f_{tT}^{av}]), \quad (1)$$

where Conv represents the convolutional network.

3.4 Memory-Assisted Metric Learning

To construct a meaningful embedding space for the audio-visual segment features, we maintain a feature bank $\boldsymbol{B}$ during training that stores the high event-related audio-visual features across different videos. We then utilize the features in $\boldsymbol{B}$ to generate prototypes of different events for event classification and constrain the relation between different segments of the same event for event localization.

3.4.1 ***Feature Consistency of Memory Bank.*** As pointed out in [6], memory bank mechanisms may encounter inconsistency issues between the rapidly changing encoder-generated features and the stored features in the bank, which were generated by the previous encoder during training. To overcome this issue, we adopt a strategy where we do not directly store the fused audio-visual feature in the bank. Instead, we store the corresponding raw audio and visual data into the raw data bank, as shown in Figure 3. We then extract fused features from these raw data using the spatial-temporal audio-visual fusion encoder and learn the embedding space using the fused features. This way, we ensure that the features in the bank are consistent with the encoder-generated features, thereby enhancing the effectiveness of the memory bank mechanism.

3.4.2 ***Bank Update Strategy.*** To make the Audio-Visual Feature Bank store high event-related audio-visual features while maintaining efficiency, we design a bank update strategy based on the Jensen-Shannon Divergence (JSD) between audio and visual features, and the bank is updated during every training batch according to the bank update strategy. Specifically, given the intermediate audio feature F_t^a and visual feature $\bar{F}_t^v$, we use two linear layers with the softmax function performed at each segment to generate the uni-modal event category prediction distributions $u^a \in \mathbb{R}^{T\times N}$ and $u^v \in \mathbb{R}^{T\times N}$, respectively. Then, we define a uni-modal event classification loss to optimize the linear layers:

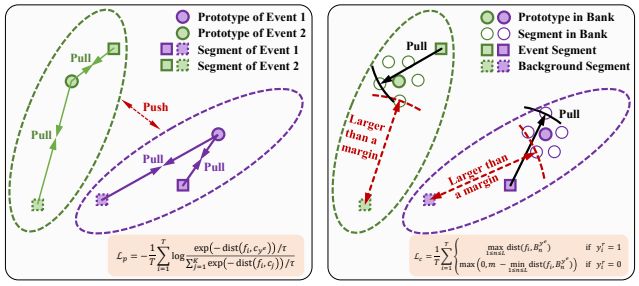
$$\mathcal{L}_{umc} = -\frac{1}{T}\sum_{i=1}^{T} y_i^r (\log u_{i,y^e}^a + \log u_{i,y^e}^v). \quad (2)$$

Next, we measure the relationship between the audio and visual features of the segment using the relation score based on the JSD and the event relevance labels. For the fully-supervised setting, we define the relation score as:

$$s_i = 1 - \text{JSD}(u_i^a, u_i^v) + y_i^r. \quad (3)$$

For the weakly-supervised setting where the event relevance labels are not available, we simply define the relation score as:

$$s_i = 1 - \text{JSD}(u_i^a, u_i^v). \quad (4)$$



(a) Event Prototypes Alignment (b) Event-Aware Localization Alignment

Figure 4: The illustration of our proposed Event Prototypes Alignments and Event-Aware Localization Alignment.

We assume that the audio-visual segments with higher relation scores are more likely to contain the same event. Thus, for each upcoming segment, we compare its relation score with the relation scores of all segments in the raw data bank that share the same event category. If the segment's relation score is higher than the relation score of any segment in the bank, we remove the segment with the lowest relation score and input the new one. In this way, the raw data bank can store and update representative audio-visual feature pairs for each event category during training.

3.4.3 ***Event Prototypes Alignment.*** After the audio-visual feature bank is obtained, we propose to use the features in the bank as the event prototypes to learn the embedding space for audio-visual features. The motivation behind this approach is that prototypes can capture the high-level representation of the events and provide better discriminative feature representations for the construction of the embedding space. Specifically, we define the prototype c_n of event n as the mean vector of the corresponding event features $\{B_1^n, \ldots, B_L^n\}$ in the feature bank $\boldsymbol{B}$:

$$c_n = \frac{1}{L}\sum_{q=1}^{L} B_q^n, \tag{5}$$

where L is the number of audio-visual features of event n in the feature bank. Then, for each audio-visual feature $f_i \in F$, we compute the distribution p over the classes based on the Euclidean distance between the feature and the prototypes in the embedding space:

$$p(y^e = n|f_i) = \frac{\exp\left(-\text{dist}(f_i, c_n)/\tau\right)}{\sum_{j=1}^{N}\exp\left(-\text{dist}(f_i, c_j)/\tau\right)}, \tag{6}$$

where $\text{dist}(\cdot)$ measures the Euclidean distance between two variables, τ is the temperature parameter that controls the range of the scores in the softmax. Finally, we define the prototype loss L_p as the negative log-likelihood of the predicted distribution over the ground-truth labels y^e:

$$\mathcal{L}_p = -\frac{1}{T}\sum_{i=1}^{T}\log p(y^e = n|f_i). \tag{7}$$

As shown in Figure 4(a), by aligning the audio-visual features with their corresponding prototypes, we can better learn the embedding space by reducing the intra-class variance and inter-class similarity, leading to improved audio-visual event recognition performance.

3.4.4 ***Event-Aware Localization Alignment.*** To model the relationship between event-related and background audio-visual features in the embedding space, we propose a loss function based on contrastive learning to constrain the feature distribution, aiming to ensure that the event-related audio-visual features are closer to the features in the feature bank than the background features.

During training, as shown in Figure 4(b), we minimize the distance between event-related segments and segments in the bank of the same event to increase their similarity in the embedding space. We also maximize the distance between background segments and segments in the bank of the same event by a margin m to increase their separation in the embedding space.

Specifically, we perform contrastive learning between the segment feature f_i of event n and the features $B^n = \{B_1^n, \ldots, B_q^n, \ldots, B_L^n\}$ of the same event in the feature bank. If f_i is an event-related segment, we minimize the largest distance between f_i and all B_q^n. Otherwise, if f_i is a background segment, we maximize the smallest distance between them by a margin m:

$$\begin{aligned}\mathcal{L}_c = \frac{1}{T}\sum_{i=1}^{T} & y_i^r \max_{1\le q\le L}\text{dist}(f_i, B_q^{y^e}) \\ & + (1-y_i^r)\max(0, m - \min_{1\le q\le L}\text{dist}(f_i, B_q^{y^e})).\end{aligned} \tag{8}$$

3.4.5 ***General Event Classification & Localization.*** Besides the event-specific localization preferences, we also expect the embedding space can capture the general localization preferences. Therefore, we also adjust the embedding space based on the general event classification and localization losses. Specifically, we first use a linear classifier and the mean pooling operation to generate the event category scores $\bar{l}^e \in \mathbb{R}^C$:

$$l_i^e = f_i W_e + b_e, \tag{9}$$

$$\bar{l}^e = \text{Mean}(\{l_1^e, \ldots, l_i^e, \ldots, l_T^e\}), \tag{10}$$

where $W_e \in \mathbb{R}^{d\times N}$ and $b_e \in \mathbb{R}^N$ are the weight and bias of the linear layer, respectively. At the training stage, we compute the cross-entropy loss given the event category scores and the video-level event category label y^e:

$$\mathcal{L}_e = -\log\frac{\exp\left(\bar{l}_{y^e}^e\right)}{\sum_{n=1}^{N}\exp\left(\bar{l}_n^e\right)}. \tag{11}$$

Then, for each segment f_i in the video, we use a linear classifier followed by the sigmoid function to generate the event relevance score $l_i^r \in \mathbb{R}$:

$$l_i^r = \text{sigmoid}(f_i W_r + b_r), \tag{12}$$

where $W_r \in \mathbb{R}^d$ and $b_e \in \mathbb{R}$ are the weight and bias of the linear layer, respectively. At the training stage, we compute the binary cross-entropy loss using relevance scores $l^r = \{l_1^r; \ldots; l_T^r\}$ and the binary segment-level event relevance label y^r:

$$\mathcal{L}_r = -\frac{1}{T}\sum_{i=1}^{T}(y_i^r \log l_i^r + (1-y_i^r)\log(1-l_i^r)). \tag{13}$$

Table 1: Comparison of our method with the existing methods under both fully-supervised and weakly-supervised settings. "VGG+VGGish" means the visual features are extracted by VGG model and the audio features are extracted by VGGish model.

Method	Fully-Supervised		Weakly-Supervised	
	VGG + VGGish	VGG + CNN14	VGG + VGGish	VGG + CNN14
AVEL (Tian et al. 2018 [22])	72.7	73.6	66.7	70.2
DAM (Wu et. al. 2019 [27])	74.5	77.3	-	74.0
AVIN (Ramaswamy et. al. 2020 [19])	75.2	-	69.4	-
AVT (Lin et. al. 2020 [13])	75.8	76.7	70.2	74.3
CMRAN (Xu et. al. 2020 [29])	77.4	78.3	73.0	74.3
MPN (Yu et. al. 2021 [31])	77.6	72.0	-	-
PSP (Zhou et. al. 2021 [34])	77.8	78.7	73.5	75.7
CMBS (Xia et. al. 2022 [28])	79.3	-	74.2	-
MM-Pyramid (Yu et. al. 2022 [32])	77.8	-	73.2	-
SPAV (Wu et. al. 2022 [26])	-	80.1	-	76.3
Ours	**80.4**	**81.3**	**77.2**	**78.7**

3.5 Optimization

3.5.1 ***Fully-Supervised Setting.*** In the fully-supervised setting, both the video-level event category label y^e and segment-level event relevance label y^r are available, so we sum all the losses together as the optimization objective function:

$$\mathcal{L}_f = \lambda_1 \mathcal{L}_e + \lambda_2 \mathcal{L}_r + \lambda_3 \mathcal{L}_{umc} + \lambda_4 \mathcal{L}_p + \lambda_5 \mathcal{L}_c, \tag{14}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are trade-off hyper-parameters.

3.5.2 ***Weakly-Supervised Setting.*** In the weakly-supervised setting, only the video-level event category label y^e is available. Following previous work [22, 28], we formulate the weakly supervised problem as multiple-instance learning (MIL) problem. Specifically, after obtaining the relevance scores $l^r = \{l_1^r; \ldots; l_T^r\}$, we aggregate the individual predictions into a video-level relevance prediction by mean pooling and compute the video-level relevance loss:

$$\bar{l}^r = \frac{1}{T} \sum_{i=1}^{T} l_i^r, \tag{15}$$

$$\mathcal{L}_r^w = -(\bar{y}^r \log \bar{l}^r + (1 - \bar{y}^r) \log(1 - \bar{l}^r)), \tag{16}$$

where $\bar{y}^r$ are the video-level relevance label and $\bar{y}^r = 1$ if the video contains any event and $\bar{y}^r = 0$ otherwise.

Moreover, we redefine the weakly-supervised version of $\mathcal{L}_{umc}$ and $\mathcal{L}_c$ as:

$$\mathcal{L}_{umc}^w = -\frac{1}{T} \sum_{i=1}^{T} (\log u_{i,y^e}^a + \log u_{iy^e}^v), \tag{17}$$

$$\begin{aligned} \mathcal{L}_c^w = \frac{1}{T} \sum_{i=1}^{T} & l_i^r \max_{1 \le q \le L} \operatorname{dist}(f_i, B_q^{y^e}) \\ & + (1 - l_i^r) \max(0, m - \min_{1 \le q \le L} \operatorname{dist}(f_i, B_q^{y^e})). \end{aligned} \tag{18}$$

To summarize, the overall objective function for the weakly-supervised setting is:

$$\mathcal{L}_w = \lambda_1^w \mathcal{L}_e + \lambda_2^w \mathcal{L}_r^w + \lambda_3^w \mathcal{L}_{umc}^w + \lambda_4^w \mathcal{L}_p + \lambda_5^w \mathcal{L}_c^w, \tag{19}$$

where $\lambda_1^w, \lambda_2^w, \lambda_3^w, \lambda_4^w$ and λ_5^w are trade-off hyper-parameters.

3.6 Inference with Event-Aware Double Branches

At the inference stage, we separately predict the video-level event category and the segment-level event relevance and combine them together to get the final prediction. Specifically, we first compute the distance between all the segments $f_i \in F$ of the video and each event prototype, and choose the event n with the smallest distance as the predicted event category:

$$n = \min_{1 \le n \le N} \frac{1}{T} \sum_{i=1}^{T} \operatorname{dist}(f_i, c_n). \tag{20}$$

Then, we compute the Euclidean distances between the segment f_i and each bank feature of the predicted event n and choose the smallest distance α_i:

$$\alpha_i = \min_{1 \le q \le L} \operatorname{dist}(f_i, B_q^n). \tag{21}$$

If α_i is smaller than the margin m, the segment f_i will be predicted as event n, otherwise, it will be predicted as background.

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 ***AVE Dataset.*** The AVE dataset is collected from AudioSet [4] by Tian et al. (2018) and contains 4,143 videos covering 28 event categories. Each video contains one event and is evenly sampled into 10 audio-visual segments. Then, the event categories are labeled for each video on the segment level. The visual and audio features are pre-extracted following the previous work [22, 34]. For visual features, we use the VGG-19 model [21] pre-trained on the ImageNet dataset [3] to extract a 512×7×7-D feature map for each visual segment. For audio features, we employ either the VGGish model [7] or the CNN14 model [11], pre-trained on AudioSet [4], to obtain a 128-D or a 2048-D feature map for each audio segment, respectively. Our experiments are conducted under two settings: the first uses VGG-19 for visual features and VGGish for audio features (i.e., **VGG + VGGish**), while the second utilizes VGG-19 for visual features and CNN14 for audio features (i.e., **VGG + CNN14**).

Table 2: Ablation Study of our model. "– Audio Modality" means removing the components related to audio encoding in our model and training without audio features. "– EADB" means removing our proposed Event-Aware Double Branches.

Setting	Fully-Supvervised		Weakly-Supervised	
	VGG + VGGish	VGG + CNN14	VGG + VGGish	VGG + CNN14
Full Model	**80.4**	**81.3**	**77.2**	**78.7**
– Audio Modality	63.7	64.1	61.1	61.2
– Visual Modality	65.3	74.1	62.5	71.3
– Transformer	79.1	80.7	75.9	77.3
– Spatial Fusion	78.3	80.8	76.1	78.2
– Temporal Fusion	75.7	79.5	74.3	76.9
– Spatial and Temporal Fusion	74.0	77.3	73.9	75.2
– JSD-based Bank Update	79.4	80.3	75.2	77.0
– Bank Feature Consistency	77.6	79.1	74.6	76.2
– EADB, + single branch	78.3	79.5	75.1	76.7
– EADB, + double branch	78.9	80.1	75.8	77.1
– EADB, + a variant of EADB	79.1	80.4	76.4	77.6

4.1.2 ***Evaluation Metrics.*** To make a fair comparison with previous methods, we employ the accuracy (*acc*) of the predicted event categories over all segments to evaluate the model performance in both fully-supervised and weakly-supervised settings.

4.1.3 ***Implementation Details.*** We implement our model using the PyTorch [18] library. We set the number of transformer blocks as 1 for all the encoders and decoders in our model. The number of attention heads, dimension of hidden states, and feed-forward layers are set to 4, 256, and 1,024 in all transformer blocks, respectively. During the training stage, we train the model for 300 epochs with a batch size of 128. We adopt the Adam optimizer [10] with an initial learning rate of 3e-4, and the learning rate is gradually decayed by the cosine annealing schedule [16]. We set the hyper-parameters τ to 1, λ_1, λ_2, λ_1^w, λ_2^w to 1, λ_4, λ_5, λ_4^w, λ_5^w to 0.1, and λ_3, λ_3^w to 0.01.

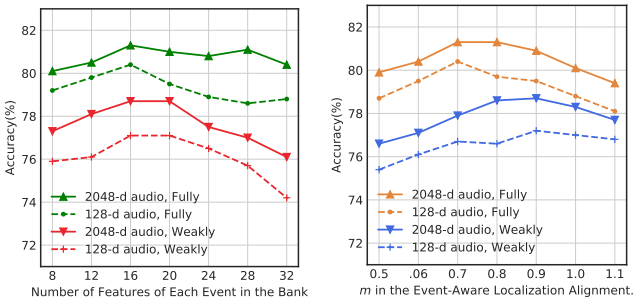
4.2 Comparison with Existing Methods

Table 1 provides a comparison of the performance of our method with state-of-the-art methods in both fully-supervised and weakly-supervised settings. In the fully-supervised setting, our method achieves state-of-the-art performance with an accuracy of 80.4% and 81.3% with 128-d and 2048-d audio feature dimensions, respectively. This indicates that our method is highly effective in utilizing fully labeled data to recognize events in videos. In the weakly-supervised setting, our method outperforms all existing methods with an accuracy of 77.2% and 78.7% with 128-d and 2048-d audio feature dimensions, respectively. This demonstrates the effectiveness of our proposed approach for recognizing events in videos when only video-level category labels are available. Furthermore, our method achieves better results with a 2048-d audio feature compared to 128-d audio feature in both fully-supervised and weakly-supervised settings. This suggests that a higher-dimensional audio feature representation can capture more discriminative audio features, leading to improved recognition performance. Overall, the experimental results demonstrate the effectiveness of our method, which outperforms state-of-the-art methods in both fully-supervised and weakly-supervised settings, with the 2048-d audio feature dimension providing better recognition performance.

4.3 Ablation Study

4.3.1 ***Effect of the Spatial-Temporal Audio-Visual Fusion Encoder.*** In the experiment results shown in Table 2, we conduct an ablation study to investigate the effect of the components of the Spatial-Temporal Audio-Visual Fusion Encoder. We first remove the model components related to audio and visual encoding, respectively, and train the model without audio or visual features. The results indicate that both modalities are crucial for the AVEL task, as removing either the audio or visual modality encoding components from the model results in significant performance drops. Next, we examine the effect of spatial and temporal fusion by removing the model components related to either spatial fusion, temporal fusion, or both of them. The results show that removing either spatial or temporal fusion from the model results in performance drops, with temporal fusion having a greater impact on performance than spatial fusion. Finally, we replace the Transformer block in our model with the vanilla attention layer [24], which is widely used in previous AVEL methods [28, 30, 34]. The results indicate that even without the Transformer block, our proposed spatial-temporal fusion model still achieves state-of-the-art results among all existing methods, validating the effectiveness of our model and learning strategies. Overall, these results demonstrate that the proposed model architecture and fusion strategies are effective for AVEL.

4.3.2 ***Effect of the Memory Bank.*** To study the effectiveness of the memory bank, we compare our memory bank strategy with the following variants: 1. Replace the JSD-based bank update strategy with choosing event-related segments randomly. 2. Replace the bank feature consistency operation with directly storing the audio-visual features. As shown in Table 2, it can be observed that replacing the JSD-based bank update component results in a drop in performance, with scores of 79.4% and 80.3% for the fully-supervised setting with 128-d and 2048-d feature dimensions, and scores of 75.1%



(a) (b)

Figure 5: (a) Impact of bank size for each event. (b) Impact of using different m in the Event-Aware Localization Alignment.

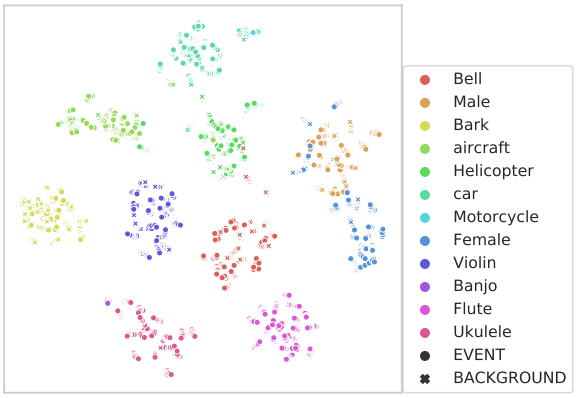


Figure 6: The t-SNE visualization of the audio-visual features in the test set. The • symbol denotes the event segments and the × symbol denotes the background segments.

and 76.7% for the weakly-supervised setting. On the other hand, replacing the bank feature consistency operation results in a more significant drop in performance. The results show the importance of the Memory Bank in MAML.

4.3.3 ***Effect of the Event-Aware Double-Branch Localization Paradigm.*** To study the effectiveness of our proposed Event-Aware Double-Branch Localization (EADB) paradigm, we compare it with the following variants: 1. Replacing EADB with a single (N+1)-class classification branch (as illustrated in Figure 1(a)); 2. Replacing EADB with the traditional double-branch (as illustrated in Figure 1(b)); and 3. Replacing EADB with a variant of EADB, which maintains the dependence between the two branches but removes the MAML and only uses non-parameter-sharing classifiers (i.e., ablating the parameter sharing among all classifiers in the two branches in Figure 1(c)). The results in Table 2 suggest that EADB plays an important role in the model's performance, and replacing it with other paradigms could have a negative impact on localization accuracy. Furthermore, among the models with different AVEL paradigms, the variant EADB achieves the highest accuracy, followed by the double-branch model and the single-branch model, which validate the effectiveness of utilizing event-specific localization preferences.

4.4 Model Analysis

4.4.1 ***Impact of Bank Size for Each Event.*** We conduct experiments to investigate the influence of different feature numbers of each event category in the memory bank. We set the number of each event category in the memory bank from 8 to 32 step by 4 and then train and test the model with different settings. Figure 5(a) shows the impact of bank size for each event on the AVEL accuracy. We can observe that the performance of the model varies with the number of features in the memory bank. When the number of features is 16, the model achieves the highest accuracy for both the 128-d and 2048-d audio inputs with fully-supervised and weakly-supervised training. However, when the number of features is too small or too large, the classification accuracy decreases. This suggests that maintaining an appropriate number of features for each event category is important for achieving good performance.

4.4.2 ***Impact of Margin m in the Event-Aware Localization Alignment.*** Furthermore, we conduct experiments to investigate the impact of the margin m in the proposed event-aware double-branch localization paradigm. We set m from 0.5 to 1.1 step by 0.1, and then train and test the model with different m. The results shown in Figure 5(b) reveal that the accuracy of the model is affected by the choice of m value. Specifically, when m is set to 0.7 for the fully-supervised setting and 0.9 for the weakly-supervised setting, the model achieves the highest accuracy among all tested values. Moreover, when m is smaller or larger, the accuracy of the model decreases. This suggests that a moderate margin value can balance the influence of the intra-event variation and inter-event distance in the embedding space, resulting in improved localization accuracy.

4.4.3 ***Visualization of Learned Embeddings.*** Figure 6 shows the visualization of the fused audio-visual features using t-SNE [23]. It can be observed that most samples are divided into different semantic clusters according to their events correctly after training. Also, the background features of each event category are closer to their e-related features than to the features of other categories. The visualization result indicates that our method can model the relation between event-specific background features and event-related features and construct the semantic embedding space effectively.

5 CONCLUSION

In this paper, we aim to perform audio-visual event localization task by leveraging event-specific localization preferences. To this end, we propose a novel event-aware double-branch localization paradigm. Moreover, we design a spatial-temporal audio-visual fusion encoder to capture the spatial and temporal interaction between audio and visual modalities, and introduce a memory-assisted metric learning framework to well align the representation space. Experimental results demonstrate the effectiveness of our proposed method in both fully-supervised and weakly-supervised settings.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grants Nos. 61972192, 62172208, 61906085, 41972111. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. 2022. MeMOT: multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8090–8100.

[2] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258* (2022).

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[5] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[8] Bo Ji and Angela Yao. 2022. Multi-Scale Memory-Based Video Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1919–1928.

[9] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. 2022. Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4350–4360.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[12] Mingxiao Li and Marie-Francine Moens. 2022. Dynamic Key-Value Memory Enhanced Multi-Step Graph Reasoning for Knowledge-Based Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10983–10992.

[13] Yan-Bo Lin and Yu-Chiang Frank Wang. 2020. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*.

[14] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1665–1673.

[15] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. 2022. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19366–19375.

[16] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[17] Tanvir Mahmud and Diana Marculescu. 2023. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5158–5167.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[19] Janani Ramaswamy. 2020. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4372–4376.

[20] Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. 2022. Dual Perspective Network for Audio-Visual Event Localization. In *European Conference on Computer Vision*. Springer, 689–704.

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Yapefng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[23] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[25] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* (2014).

[26] Yiling Wu, Xinfeng Zhang, Yaowei Wang, and Qingming Huang. 2022. Span-based Audio-Visual Localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1252–1260.

[27] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. 2019. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6292–6300.

[28] Yan Xia and Zhou Zhao. 2022. Cross-Modal Background Suppression for Audio-Visual Event Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19989–19998.

[29] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. 2020. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3893–3901.

[30] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. 2020. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 279–286.

[31] Jiashuo Yu, Ying Cheng, and Rui Feng. 2021. Mpn: Multimodal parallel network for audio-visual event localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[32] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. 2022. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6241–6249.

[33] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic memory networks for short text classification. *arXiv preprint arXiv:1809.03664* (2018).

[34] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. 2021. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8436–8444.